

# Missing Data Imputation Over Academic Records Of Electrical Engineering Students

Esteban Jove<sup>a</sup>, Patricia Blanco-Rodríguez<sup>b</sup>, José-Luis Casteleiro-Roca<sup>a</sup>, Héctor Quintián<sup>a</sup>,  
Francisco Javier Moreno Arboleda<sup>c</sup>, José Antonio López-Vázquez<sup>a</sup>, Benigno Antonio  
Rodríguez-Gómez<sup>a</sup>, María del Carmen Meizoso-López<sup>a</sup>, Andrés Piñón-Pazos<sup>a</sup>, Francisco Javier  
de Cos Juez<sup>d</sup>, Sung-Bae Cho<sup>e</sup> and José Luis Calvo-Rolle<sup>a</sup>

<sup>a</sup> *University of A Coruña, Department of Industrial Engineering, Avda. 19 de febrero s/n, 15405, Ferrol, A Coruña, Spain*

<sup>b</sup> *University of Oviedo, Department of Construction and Manufacturing Engineering, Campus de Viesques, 33204, Gijón, Spain*

<sup>c</sup> *Universidad Nacional de Colombia, Facultad de Minas, Sede Medellín*

<sup>d</sup> *University of Oviedo, Department of Mining Exploitation, Calle San Francisco, 1, 33004, Oviedo, Spain*

<sup>e</sup> *Yonsei University, Department of Computer Science, Seoul 03722, South Korea*

**Abstract:** Nowadays, the quality standards of higher education institutions pay special attention to the performance and evaluation of the students. Then, having a complete academic record of each student, such as number of attempts, average grade and so on, plays a key role. In this context, the existence of missing data, that can happen for different reasons, leads to affect adversely interesting future analysis. Therefore, the use of imputation techniques are presented as an helpful tool to estimate the value of missing data. This works deals with the academic records of engineering students, in which imputation techniques are applied. More specifically, it is assess and compared the performance of the Multivariate Imputation by Chained Equations methodology (MICE), the Adaptive Assignment Algorithm (AAA) based on Multivariate Adaptive Regression Splines (MARS) and a hybridisation based on Self-Organisation Maps (SOM) with Mahalanobis distances and AAA algorithm. The results show that proposed methods obtain successfully results regardless the number of missing values, in general terms.

KEY WORDS Student Performance; Data imputation; SOM; MARS; MICE; AAA

*Received November 5, 2021*

## 1 Introduction

According to guidance of quality assurance systems under the European Higher Education Area (EHEA), studies tracking is not only legally regulated but it is also compulsory for official university degrees [?]. Consequently, internal quality systems of educational institutions make great efforts to enhance their quality ratios or indicators in terms of academic results and performance [?]. Faculties or higher education schools need thus tools to support or assist in this complicated task [? ? ? ?].

In this context, missing data on academic records implies a significant problem that must be tackled. Then, the use of techniques to estimate missing values are presented as a good solution.

In earlier research works, common methodologies involved obtaining a model based on an historical dataset, through either traditional techniques or more advanced procedures [? ? ? ? ? ? ? ?]. This traditional method could be a problem in general terms, given the need to have previous cases with similar performance [? ? ? ? ? ? ? ? ? ?].

It must be emphasised that the case under study could change, so the model must be adaptive for novel cases with different casuistic and performance [? ? ? ? ? ?]. Therefore, imputation procedures based on evolutionary methods appear to be a good solution to accomplish the studies tracking task.

This paper evaluates different imputation techniques to estimate missing values on academic records of university students. In particular, the initial dataset is composed by the grades and number of attempts of electrical engineering degree students. The students academic records are from the Polytechnic University College (in the University of A Coruña). The dataset includes a total of 7 academic course, from 2001/2002 to 2008/2009.

The techniques employed are the Adaptive Assignment Algorithm (AAA) [? ? ?], the Multivariate Imputation by Chained Equations (MICE) algorithm [?] and a hybridisation based on Self-Organization Maps (SOM) with Mahalanobis distances and AAA algorithm, which is a combination of pattern recognition and

\*Correspondence to: Corresponding author address. E-mail: jose.luis.casteleiro@udc.es

machine learning methodologies. The performance of each one on the filling in missing data task is assessed and compared.

The results obtained over a real dataset show that, in general terms, for one and two missing values, the AAA algorithm offers better performance than MICE. On the other hand, with three missing values, the MICE algorithm achieve better results. Proper combination of these two algorithms appears to be a potential good solution that requires to clearly establish the boundary between application of both methodologies. Furthermore, the hybridisation based on SOM, Malahanobis distances and AAA offers successfully results when it is applied to a dataset with a number of missing data randomly selected.

This paper is structured as follows: after the present introduction, the case of study is detailed in the next section. Then, the data imputation techniques used, and the approach are explained. The outcomes and the best model are exposed in the results section. Finally, the conclusions and future works are exposed.

## 2 Case of study

The present work deals the attempts prediction by missing data imputation of an engineering degree. To achieve this goal, a dataset consisting on the academic records of electrical engineering degree students was used. This degree is taught at the Polytechnic University College (EUP) of the University of A Coruña (UDC) and it is structured in three courses with a total amount of 25 subjects: the first and second courses have 9 subjects and the last course has 7 subjects. The set of subjects for each degree corresponds to 236 academic credits.

To achieve the final dataset used on this research, two different steps were followed. At first, the academic data of students enrolled in the EUP from 1996/1997 to 2008/2009 were considered. With this data, an initial statistical analysis was performed taking into account different factors such as access method, admission grade, average grade, gender or geographic origin. From this data, some conclusions are reached:

- The 80 % of the students are male.
- The 99.4 % of the students come from Galicia.
- The 91.65 % of the enrolled people come from secondary education.
- The first enrolment is done at an average age of 20.86.
- The students obtain their graduate at an average age of 26.40.
- An average of 5.96 courses are needed to finish the studies.

After this step, two different groups of students were discarded: the students who enrolled the university before 2001/2002 due to lack of information about their academic records, and the students that abandoned their studies in the EUP. The final dataset considered on this work includes for each student the number of attempts to pass each subject and the grade achieved on each one.

The dataset includes the next information:

- Number of attempts to pass each subject.
- The grade achieved for each subject.
- Academic background: secondary school or vocational education and training.
- Admission grade.

With the aim of checking the performance of the used algorithms, different percentages of missing values were emulated. This step is vital to establish the proper boundaries of each method and hence, develop a hybrid model useful in a wider range of cases.

## 3 Data imputation techniques

This section describes in a detailed way the data imputation methods used in the present research work.

### 3.1 MICE algorithm

The MICE algorithm was developed by van Buuren and Groothuis-Oudshoorn [? ]. It is based on Markov Chain Monte Carlo Method, where the state space is the dataset of all the values imputed.

Like other Markov Chain methods, given the fact that the results must converge, this algorithm needs to meet the following characteristics [? ? ? ? ]:

- *Irreducible*: Chain has to able to comprise all elements of the space of state.
- *Aperiodic*: Chain must not fluctuate among different states.

- *Recurrence*: Any chain of this type could be considered recurrent if the likelihood of the Markov chain to start from  $i$  and return to  $i$  is equal to the unit.

From a practical point of view, the MICE algorithm convergence is obtained with a very low quantity of iterations, generally from 5 to 20 [? ? ? ? ?].

In accordance with the know-how of the method owner, five iterations are usually adequate. However, some different contingencies need more iterations.

Then, the iterations quantity for the present research work was fixed to five, a very low value compared with other Markov Chain Monte Carlo Methods applications, where it is necessary a large quantity of iterations, even thousands. It is very important to emphasise that each iteration of the MICE method could spend within some minutes up to hours. Also, the time spent on the iteration usually depends on the number of variables implied, and it does not depend so much on the amount of cases. Furthermore, it must be taken into account that the data imputed could have an appreciable random noise, that depends on the variables correlations.

If the variables are independent or if the correlations between them are low, then the algorithm convergence tends to be faster. It is also remarkable that high rates of missing data (more than 20 %) lead to slow convergence problems.

The necessary steps for the multivariate missing data imputation are the following [? ]:

1. Specify an imputation model  $P(Y_j^{mis}|Y_j^{obs}, Y_{-j}, R)$  for variable  $Y_j$  with  $j = 1, \dots, p$   
MICE algorithm gives the later  $R$  distribution using interactive sampling from the represented conditional formula shown above. The  $R$  parameters are specific to the respective conditional densities and are not necessarily the product of a factorisation of the true joint distribution.
2. For each  $j$ , fill in starting imputations  $Y_j^0$  by random draws from  $Y_j^{obs}$
3. Repeat for  $t = 1, \dots, T$  (iterations)
4. Repeat for  $j = 1, \dots, p$  (variables)
5. Define  $Y_{-j}^t = (Y_1^t, \dots, Y_{j-1}^t, Y_{j+1}^{t-1}, \dots, Y_p^{t-1})$  as the currently complete data except  $Y_j$
6. Draw  $\phi_j^t \sim P(\phi_j^t|Y_j^{obs}, Y_{-j}^t, R)$
7. Draw imputations  $Y_j^t \sim P(Y_j^{mis}|Y_j^{obs}, Y_{-j}^t, R, \phi_j^t)$
8. End repeat  $j$
9. End repeat  $t$

$Y$  represents a  $n \times p$  matrix of partially-observed sample data,  $R$  is a  $n \times p$  matrix, 0 – 1 response indicators of  $Y$ , and  $\phi$  represents the parameters space. Please note that with MICE imputation [? ], the initial guesses for all missing elements are given for the  $n \times p$  matrix of partially observed sample.

The dataset is divided into two subsets for each variable with missing elements, one of them containing all the missing data.

If the subset with all the data available is in regression with all other variables, the subset missing is predicted from the regression, by the replacement of the missing values by those achieved from the regression. This methodology is repeated for all variables with missing elements.

To conclude the procedure, all missing elements are imputed according to the above explained algorithm and the regression and predictions are repeated till the stop criterion is reached or, in this case, till some quantity of successive iterations are within the defined tolerance for each of the values imputed.

### 3.2 The Adaptive Assignment Algorithm (AAA) algorithm

With the aim to explain the AAA method [? ], it is necessary to assume that the dataset includes a set of variables  $v_1, v_2, \dots, v_n$ . To obtain the missing values of the  $i$  –  $th$  column all rows without missing values in that column are used and, then, some MARS models are estimated.

It is possible to find rows with a very variable quantity of missing data, from 0 (no missing values) to  $n$  (all values are missing).

Also, the columns that contains all missing values can be removed so that they will not be used for the model calculation and will not be imputed. Furthermore, a number of missing data between 0 and  $n - 2$  is feasible (all variables except one with missing values).

Then, the number of MARS models that must be taken into account to estimate the missing values in the column  $v_i$  of a dataset formed by  $n$  variables  $v_1, v_2, \dots, v_n$  is shown in equation 1.

$$\sum_{k=1}^{n-1} \binom{n-1}{k} \quad (1)$$

The present research work has 10 variables, so the maximum number of MARS models is 5,110 (511 by variable). Once all the available models were obtained, the next step is to compute all the missing data of each row. To achieve this goal, the models without missing data at this row must be used. When the model was not obtained, missing data must be substituted by the column average value. Remark that the large datasets with not too high missing data percentage are infrequent.

In general, with the use of this algorithm, if a value can be estimated with more than one MARS model, then, it must be checked the model with the highest quantity of input variables; the value could be obtained with some of those chosen models.

Exceptionally, if there is not available model for estimation, the variable average value can be employed in the imputation process.

### 3.3 The hybrid proposed methodology

The algorithm described on this section is based on the hybridisation of the Self-Organized Maps neural networks with the Mahalanobis distances. Also, the hybrid topology achieved is combined with the above shown algorithm AAA [? ], that is based on Multivariate Adaptive Regression Splines. The new proposed methodology performance is better than the developed on previous research when it was used with the same dataset. It is considered a hybridisation due to the fact that this proposal combines pattern recognition and machine learning methods, achieving a hybrid model that allows the missing data imputation [? ].

If there is a dataset composed by  $c$  different variables  $v_1, v_2, \dots, v_c$  that are the matrix columns with a total number of  $r$  rows, the algorithm could be applied by following the steps described in the next subsections.

#### 3.3.1 New matrix creation of missing values as from the original data set

If the algorithm is applied to a dataset in which the missing data are going to be imputed, the present step of the algorithm is not necessary. However, in this research it is necessary to validate the proposal with a complete dataset.

$A$  is the original matrix ( $r \times c$ ) of  $r$  rows and  $c$  columns. With the aim to obtain a matrix with missing data, a rate of  $p$  elements must be removed.  $B$  will be the new ( $r \times c$ ) matrix, with a rate  $p$  of missing elements.

The removal is accomplished in a randomly manner, then, the imputation method that is going to be checked is defined as Missing Completely At Random (MCAR).

#### 3.3.2 The reduced matrix creation

It is created a new matrix where all the rows that have missing data are removed. The new matrix is denominated as  $B^{red}$ . Despite the quantity of rows  $s$  ( $s \leq r$ ) of this matrix will change depending on the matrix to be imputed, in cases where the data removal was made completely at random, and proportionally to  $p$ , the quantity of remaining rows  $u$  follows the equation 2.

$$u = r \cdot (1 - p)^c \quad (2)$$

Where:

- $p$ : proportion of missing data taken into account.
- $r$ : number of rows of the original matrix.
- $c$ : number of columns of the data matrix.

Then, the  $B^{red}$  matrix is normalised.

#### 3.3.3 The director vectors obtaining through Self-Organized Maps

The Self-Organized Maps is a kind of unsupervised artificial neural network with a wide range of applications. One of these applications is the visualisation and the subsequent interpretation of datasets, usually high-dimensional [? ].

These kind of artificial neural networks are employed to perform the available observations, with very high accuracy, thanks to a model quantity reduction. This is why this method has been selected to accomplish this work.

$N$  is the dimension of the  $n$  director vectors,  $X(t) \in \mathbb{R}_n, t = 1, 2, \dots, n$ . Each sample vector is labelled.

All the nodes are used as a codebook vector  $Wk$  of dimension  $N$ .

The weight vectors values are obtained following the next algorithm [? ]. It is necessary to establish a number of iterations, following these next steps on each one:

1. Choose one sample vector  $X(t)$  at random.

2. Search for the nearest weight vector:  $W_c : \|X - W_c\| = \min_j \|X - W_j\|$
3. Update the weights  $W_i$  by means of the the rule in equation 3.

$$W_i(t+1) = W_i(t) + h_{ci}(t) \cdot [X(t) - W_i(t)] \quad (3)$$

Where  $h_{ci}(t)$  is the neighbour function, which, in the case of this work and being very usual in the bibliography [? ], is a Gaussian equation (equation 4).

$$h_{ci}(t) = \alpha(t) \cdot \exp\left(\frac{\|W_c - W_i\|}{2 \cdot \sigma^2(t)}\right) \quad (4)$$

The neurons weights that are at the neighbourhood  $h_{ci}(t)$  of the winning neuron are displaced near to  $X(t)$ . When the number of iterations increase, the learning rate  $\alpha(t) \in [0, 1]$  decreases monotonically. After some iterations, and if the reduction of  $\alpha(t)$  and  $\sigma(t)$  is slow, only a single node is covered by a neighbourhood, and consequently the map is created. It is necessary to remark that, if these neurons weights are close in the parameter space  $W$ , then, they are also close at the mesh.

Then, the achieved director vectors are denormalized. The chosen quantity of director vectors to implement the SOM for the present algorithm is proportional to the rows quantity in the  $B^{red}$  matrix.  $u$  is the quantity of rows of the matrix  $B^{red}$ ; the number of director vectors will be a range of values  $d = e \cdot \frac{u}{\epsilon} \in [0.05, 0.8]$ . In the results section, this range values are explained and justified.

### 3.3.4 The closest director vectors obtaining with Mahalanobis distances

This technique is a non-Euclidean distance measure based on correlations between variables [? ]. Some patterns are identified and analysed thanks to these correlations. This indicator is a very helpful way to achieve similarities over an unknown sample set to a known set. It is used on this work to compare each one of the rows of the matrix with missing data with all the director vectors.

It is defined with the equation 5.

$$d_A(x_1, x_2) = \sqrt{(x_1 - x_2)^T \cdot A \cdot (x_1 - x_2)} \quad (5)$$

$x_1$  and  $x_2$  are the sets of variables of two rows of the data matrix.  $A \in \mathbb{R}^{n \times n}$  is a positively semi-definite matrix, which is the inverse of the covariance matrix of class  $\{I\}$ . With the eigenvalue decomposition,  $A$  can be decomposed into  $A = W \cdot W^T$ .

For the present approach, the Mahalanobis distance of each vector row is calculated with at least two missing data points to all the director vectors. Remark that all those variables with missing data in the row that come from the data matrix are removed in the director vector. The director vector with the lowest Mahalanobis distance value is chosen and the missing variables are filled with the present values in the corresponding row of the director vector.

The original matrix is reconstructed and the missing data value of those rows with one or two missing data points are imputed with the AAA algorithm [? ? ? ].

## 4 Results

To check the performance of the algorithms, different percentages of missing data were tested using the validation measurements explain in the following sub-section.

### 4.1 Models validation

It has been employed cross-validation leave-one-out to analyse the interpolated data spatial error [? ? ]. The use of eight of the nine stations in the model is made to achieve the estimated value at the ninth, with the aim to obtain this station. The procedure is calculated once for each station. Then, nine times for all the process.

The three methods have been compared in terms of performance, thanks to useful common statistics indicators Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Median Absolute Deviation (MAE).

$$RMSE = \sum_{i=1}^n \sqrt{\frac{1}{n} (\hat{G}_i - G_i)^2} \quad (6)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{G}_i - G_i| \quad (7)$$

$$MAD = \text{median}(|G_i - \text{median}(G)|) \quad (8)$$

$G_i$  and  $\hat{G}_i$  are the measurements and the model-estimated respectively, and  $n$  is the number of data points of the validation set. RMSE is considered a very important model validation metric and it weights large estimation errors more strongly than small estimation errors. Furthermore, MAE is a useful complement of the measured-modelled scatter plot near the 1-to-1 line [? ? ?].

## 4.2 Validation results

It should be noted that the results shown in the next tables are obtained taking into account only ten columns of the dataset. The content of each column represents a different subject randomly selected. The percentage of missing data represents a 10% of the whole dataset. However, the number of real missing data varied depending on the test from one to three.

The results obtained with only one missing value in each case are shown on Table 1. The AAA algorithm offers better performance than MICE in this case.

The performance of each technique for two missing values is represented in Table 2. For this test, the AAA algorithm shows again better performance than the MICE. However, both performances are decreased with respect to the previous case.

If the number of missing values is raised to 3, MICE algorithm offers better results, as shown in Table 3.

Table 1: Algorithm results with 1 missing value

	MSE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Column 2	0.4333	0.0000	0.1482	0.0000	0.7992	0.0000
Column 3	0.1666	0.0000	0.0000	0.0000	0.3125	0.0000
Column 4	0.1333	0.0000	0.1482	0.0000	0.2230	0.0000
Column 5	0.2333	0.0000	0.0000	0.0000	0.4280	0.0000
Column 6	0.4000	0.0000	0.1482	0.0000	0.5908	0.0000
Column 7	0.1333	0.0014	0.0000	0.0018	0.2787	0.0018
Column 8	0.1666	0.0000	0.0000	0.0000	0.3604	0.0000
Column 9	0.0666	0.0000	0.0000	0.0000	0.1154	0.0000
Column 10	0.2000	0.0000	0.1482	0.0000	0.1087	0.0000

Table 2: Algorithm results with 2 missing values

	MSE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.0502	0.0000	0.0474	0.0001	0.0370	0.0000
Column 2	0.4766	0.0000	0.2179	0.0001	0.8325	0.0000
Column 3	0.2132	0.0001	0.0294	0.0001	0.3672	0.0001
Column 4	0.1848	0.0001	0.1769	0.0001	0.2562	0.0000
Column 5	0.2721	0.0000	0.0594	0.0001	0.4879	0.0001
Column 6	0.4707	0.0000	0.1877	0.0001	0.6390	0.0000
Column 7	0.1879	0.0015	0.0693	0.0019	0.3144	0.0019
Column 8	0.2083	0.0000	0.0435	0.0001	0.3973	0.0000
Column 9	0.1258	0.0000	0.0604	0.0000	0.1515	0.0001
Column 10	0.2630	0.0000	0.2144	0.0000	0.1614	0.0001

To take advantage of the best algorithm for each situation, a "hybrid combination" of both techniques can be taken into account. Table 4 shows the performance of the hybrid combination, where the percentage of missing values is fixed to 10% and the number of missing values is randomly selected. If the number of missing values is less than 3, the "hybrid system" uses AAA algorithm while MICE is chosen otherwise. Using this configuration, the error obtained is smaller than the MICE algorithm in all cases.

Table 3: Algorithm results with 3 missing values

	MSE		MAD		RMSE	
	MICE	AAA	MICE	AAA	MICE	AAA
Column 1	0.1629	0.8613	0.1374	0.7392	0.1462	1.0162
Column 2	0.5811	0.8746	0.3514	0.7610	0.9278	0.8031
Column 3	0.3389	0.7967	0.1134	0.9031	0.4530	1.0407
Column 4	0.2830	0.7968	0.2748	0.9244	0.3656	0.8192
Column 5	0.3680	0.6989	0.1785	1.0166	0.5548	0.7210
Column 6	0.5255	0.8510	0.3003	0.9616	0.7245	0.9048
Column 7	0.3074	0.8079	0.1905	0.7252	0.3805	0.7783
Column 8	0.3333	0.8661	0.1320	0.8535	0.4980	0.8040
Column 9	0.2277	0.8602	0.1448	1.0770	0.2629	0.8335
Column 10	0.3417	0.9510	0.3103	0.7686	0.2387	0.9942

Table 4: Algorithm results with random missing values and hybrid combination

	MSE		MAD		RMSE	
	MICE	Hybrid	MICE	Hybrid	MICE	Hybrid
Column 1	0.0545	0.0059	0.0532	0.0065	0.0430	0.0049
Column 2	0.4819	0.0045	0.2229	0.0054	0.8389	0.0036
Column 3	0.2191	0.0059	0.0350	0.0028	0.3743	0.0060
Column 4	0.1905	0.0038	0.1824	0.0050	0.2630	0.0063
Column 5	0.2783	0.0051	0.0654	0.0039	0.4925	0.0079
Column 6	0.4759	0.0047	0.1929	0.0047	0.6459	0.0045
Column 7	0.1930	0.0050	0.0747	0.0082	0.3192	0.0073
Column 8	0.2120	0.0053	0.0481	0.0037	0.4018	0.0050
Column 9	0.1312	0.0033	0.0654	0.0063	0.1580	0.0051
Column 10	0.2665	0.0028	0.2199	0.0040	0.1649	0.0016

Finally, the results obtained with the hybrid approach proposed on section 3.3 are shown in Table 5. This configuration combines the unsupervised algorithm SOM with techniques that measure similarity between two random multidimensional variables and the imputation algorithm AAA.

Table 5: Algorithm results with random missing values and the new proposed hybrid combination based in SOM, Mahalanobis distances and AAA

	MSE		MAD		RMSE	
	MICE	MAAA	MICE	MAAA	MICE	MAAA
Column 1	0.0729	0.0042	0.0732	0.0053	0.0575	0.0030
Column 2	0.5017	0.0029	0.2379	0.0041	0.8493	0.0024
Column 3	0.2380	0.0046	0.0452	0.0009	0.3936	0.0048
Column 4	0.2045	0.0022	0.1956	0.0037	0.2827	0.0044
Column 5	0.2905	0.0037	0.0757	0.0028	0.5045	0.0060
Column 6	0.4936	0.0032	0.2064	0.0035	0.6653	0.0032
Column 7	0.2108	0.0039	0.0875	0.0072	0.3319	0.0056
Column 8	0.2288	0.0040	0.0675	0.0021	0.4166	0.0035
Column 9	0.1473	0.0015	0.0803	0.0046	0.1753	0.0037
Column 10	0.2841	0.0011	0.2312	0.0022	0.1830	0.0015

Figure 1 represents the RMSE values of each column for the two best algorithms when the number of missing data varies randomly. It must be emphasised that MAAA algorithms outperforms the hybrid model, which is a combination of AAA and MICE algorithms

## 5 Conclusions and future works

Very good results have been obtained in general terms with the data imputation techniques employed in this study. It is possible to predict the students' attempts to pass subjects in the cases discussed in this paper, assuming that the data do not exist and comparing the estimate results with the real dataset.

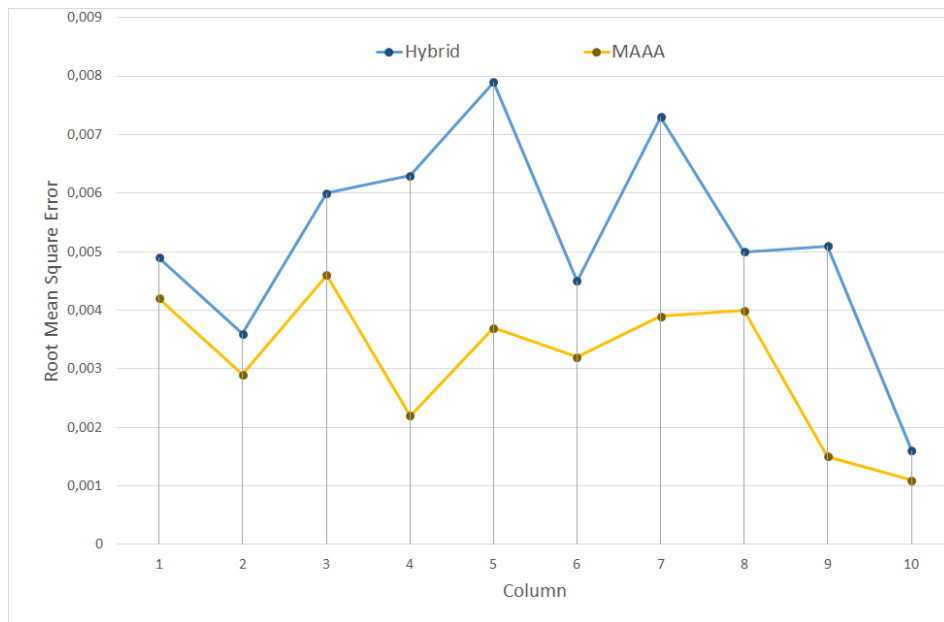


Figure 1: Distribution of the RMSE depending on the column and the algorithm

The AAA algorithm presents a very good performance for 1 and 2 missing values, with a RMSE value from 0.0000 to 0.0019. However, when the number of missing values increase to three, the RMSE varies from 0.7210 to 1.0407. On the other hand, the use of MICE algorithm offers better results with respect to AAA when three values are missing (RMSE from 0.1462 to 0.9278). This circumstance led to use a hybrid system that combines AAA and MICE algorithms depending on the number of missing values when this number is random. This hybrid configuration, outperform MICE algorithm in all cases, with a RMSE from 0.0036 to 0.0079. Furthermore, the use of MAAA leads to even better results when the number of missing values is randomly assigned between one and three. In this case, the RMSE values vary from 0.0024 to 0.0060.

These techniques could be used to predict lack of data and to estimate students' performance and hence, this would help to improve the academic records tracking and the quality standards of higher education institutions.

As future works, the use of support vector machines (SVM) [??] and other hybrid methods [??] will be considered by the authors in order to find a new algorithm with even higher performance. Also, as future works, the authors will check the procedure in other academic degrees, to ensure the generalisation of the method, and to study its application over some degrees at the University of a Coruña.

## 6 Acknowledgments

Authors greatly appreciate the support both from Spanish Ministry of Economy and Competitiveness through grant AYA2014-57648-P, and from regional Ministry of Economy and Employment through grant FC-15-GRUPIN14-017.

This work was also supported by grant funded by 2019 IT promotion fund (Development of AI based Precision Medicine Emergency System) of the Korea government (Ministry of Science and ICT).